DOCUMENT RESUME

ED 410 317 TM 027 827

AUTHOR Rudner, Lawrence M.
TITLE Early Childhood Program Research and Evaluation. ERIC/AE Digest.
INSTITUTION ERIC Clearinghouse on Assessment and Evaluation, Washington, DC.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO EDO-TM-96-10
PUB DATE Dec 96
NOTE 4p.
CONTRACT RR93002002
AVAILABLE FROM ERIC Clearinghouse on Assessment and Evaluation, 210 O'Boyle Hall, The Catholic University of America, Washington, DC 20064; toll free telephone: 800-464-3742.
PUB TYPE ERIC Publications (071)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Child Development; Cognitive Development; *Early Childhood Education; *Educational Research; Effect Size; Evaluation Methods; Intelligence; Intelligence Tests; *Program Evaluation; Test Reliability; Test Validity; *Young Children
IDENTIFIERS ERIC Digests

ABSTRACT

In educational research and evaluation, a sample of subjects usually received some type of programmatic treatment. Outcome scores for these students are then compared with outcome scores of a control or comparison group. M. Lewis and H. McGurk (1972) have pointed out that there are some implicit assumptions when this approach is applied to programs for infants and toddlers. These are: that infant intelligence is a general unitary capacity; that mental development can be enhanced by enriching the infant's experience in a few areas; and that infant scales can reflect the improvement that results from a specific enrichment experience. The contrasting view is that infants and toddlers undergo rapid, nonlinear growth and change along many lines of development. This digest examines these contrasting assumptions, first by examining the short-term and long-term consistency of test scores. Research has shown that test-retest reliability is extremely low for infants and increases moderately for toddlers, a finding that is consistent with the view of the child going through nonlinear growth. Classic studies of mental growth in infants and toddlers show inconsistent and unpredictable growth rates and demonstrate little predictive validity for early studies of mental growth. A close look at statistical implications suggests that it is not likely that the researcher or evaluator could find significant differences if they do exist. For these reasons, in lieu of control-comparison group testing, the use of case studies, the computation of effect sizes, and the examination of growth curves are suggested for evaluations of programs for infants and toddlers. (Contains 13 references.) (SLD)

ED 410 317

Early Childhood Program Research and Evaluation

Lawrence M. Rudner

ERIC/AE Digest Series EDO-TM-96-10

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM027827

BEST COPY AVAILABLE

ERIC®/AE Digest Series EDO-TM-96-10
The Catholic University of America

December 1996
Department of Education

Early Childhood Program Research and Evaluation

Lawrence M. Rudner
Catholic University of America

In research and evaluation, a sample of subjects typically receives some form of programmatic treatment then outcome scores for these students are compared with outcome scores of a control or comparison group. Lewis and McGurk (1972) point out some of the implicit assumptions when this design is applied to programs for infants and toddlers: 1) "infant intelligence is a general unitary capacity", 2) "mental development can be enhanced by enriching the infant's experience in a few specific areas", and 3) infant scales can "reflect any improvement in competence that results from a specific enrichment experience." The traditional control group-comparison group design adopts the viewpoint that frequency and nature of observable cognitive activities increase at a steady rate as a result of the growth process.

The contrasting viewpoint is that infants and toddlers are going though a period of rapid, non-linear growth and change along many interwoven lines of development (Horner, 1980). Accordingly, different levels and kinds of cognitive development would be presented by different individuals during different stages of development, short-term consistency of individual traits would be low, traits measured during infancy would have low correlations with later skills, broad programmatic treatment effects will be small, and a different research and evaluation paradigm is needed.

This digest examines these contrasting assumptions. We start by examining the short and long term consistency of test scores. We then relate this consistency to the literature on the cognitive development of infants and toddlers. We then identify gains associated with some particularly effective programs for infants and toddlers and the statistical implications of those gains. We end with a set of recommendations for the design of research and evaluation studies.

Short term consistency

Different types of reliability estimates are used to estimate the contributions of different sources of measurement error. Inter-rater reliability coefficients provide estimates of errors due to inconsistencies in judgment between raters. Estimates of internal consistency (Cronbach's alpha, Kuder Richardson formula 20 and 21) account for error due to content sampling, usually the largest single component of measurement error when testing older children and adults. Of primary interest with infants and toddlers is test-retest reliability which measures the consistency of the trait for groups of individuals.

Test-retest reliability tends to be quite low when scales are administered to infants. As the child gets older, test-retest reliabilities tend to improve. Werner and Bayley (1966) summarized studies examining the test-retest reliability of various infant measures and noted wide variations in scale scores. One study, for example, found 1 day test-retest reliabilities on the Buhrer Baby tests to range from .40 to .96 depending on the age of the infants. Another study found 2 day test-retest reliabilites on the Linfert-Hierholzer scales for 1- 2- and 3- month olds to be -.24, .44 and .69 respectively. Horner (1980) found 4-10 day test-retest reliabilities on the Bayley for 9 month old females, 9 month old males, 15 month old females and 15 month old males to be .42, .67, .96, and .76 respectively. Werner and Bayley (1966) found the percentage of agreement across two administrations of the Bayley to 8 month olds varied from 41% to 95% with a mean of 76%. With 9 and 16 month olds, Horner (1980) found slightly higher consistencies on the same items, with means of 85% for both age groups.

Thus, test-retest reliability is extremely low for infants and increases moderately for toddlers. The lack of test-retest reliability is consistent with the view of the child going through non-linear growth. It is inconsistent with the notion that the cognitive activity in infants increases at a steady rate as a result of growth.

Long Term Consistency

The classic studies of mental growth in normal infants and toddlers show inconsistent and unpredictable growth rates of these observable skills and traits. Bayley, for example, reported correlations between -.04 and .09 between scores during the first 3 months of life and scores at 18 to 36 months. Looking at race and gender with a sizeable sample, Goffeney, Henderson and Butler (1971) later found virtual no correlation between 8 month and 7 year measures. Escalona and Moriarty (1961) found virtually no correlation between 20 month and 6 to 9 year scores.

"The findings of these early studies of mental growth of infants has been repeated sufficiently often so that it is now well established that test scores earned in the first year or two have relatively little predictive validity" (Bayley, 1970). Comprehensive reviews of the literature by Stott and Ball (1965), and Thomas (1970) fully support Bayley's view and draw the same conclusion. There are notable exceptions, however. Many developmental inventories are excellent screening devices capable of identifying students with permanent cognitive disabilities.

Early childhood development

Bayley (1958) outlines the skills and behaviors that we can observe during the first years. In the early months of life, we can only observe variations in sensory-motor coordination and simple adaptive responses. These adaptive responses develop into rudimentary forms of interpersonal communication in the form of gestures, vocalizations, and basic emotional responses. Then we have language gradually developing. At first, language is tied to the immediate and concrete; later, it becomes more symbolic. The child begins to abstract and generalize his experience.

Through factor analysis, Bayley (1955) identified three distinct kinds of intellectual behaviors: sensory motor which is dominant during the first year, persistence which tends to be dominant during

the second and third year, and a general intelligence which is dominant at age 4 and the only operating factor after age 6. This third, general intelligence factor of Bayley appears to be the stable intelligence factor discussed by Binet and Cyril Burt.

The important consideration for research and evaluation is that there is no continuity across these three developmental stages. Rather, infants and toddlers develop a composite of skills that are not necessarily covariant. Scores obtained when a child is in one stage of development will be uncorrelated to scores obtained when the child is in different stage.

Statistical implications

We now turn to a key statistical consideration of the control group-comparison group model. Are our statistical tools powerful enough to detect differences when they do exist?

The power of a statistical test refers to the probability that it will lead to the rejection of a null hypothesis given that there is indeed a difference in the population (Cohen, 1988). Power depends on three parameters: the significance criterion, the reliability of the sample results, and the effect size. The significance level is usually set at $\alpha = .05$ or .01. The reliability of the sample results is a function of sample size and the reliability of the chosen measure. The effect size is the degree to which the phenomenon exists and is typically expressed as the standardized difference between group means. Given sample size, alpha level, and expected effect size, we can compute the probably of finding a significant difference in our samples.

Critical in the analysis of power is the expected effect size. How much of an effect can we expect of quality early intervention programs? Ottenbacher (1992), examined 237 effect sizes from 59 such studies. Applegate (1986) conducted a meta-analysis of 114 effect sizes from thirteen studies. One large Head Start evaluation (ACYF, 1983) coded 71 studies to look at 148 comparisons and 449 effect sizes. Depending on the age and the variables being considered, the typical effect size for infants and toddlers appears to be about .25 standard deviations.

With an effect size of .25, an alpha of .05, and a sample size of 100 subjects per group, the power of a t-test is .35. That is, there is only a 35% chance of finding an existing significant difference. If we take into account that many of the measures only have a reliability coefficient of .7, the odds of finding a significant difference drop to 28%.

Thus, the researcher or evaluator is not likely to find significant differences even when they do exist. Further, in light of the lack of long term consistency, significant differences are of little practical value.

Recommendations

We fully concur with Lewis and McGurk (1972) who wrote in their classic *Science* article that infant development scales "are unsuitable instruments for assessing the effects of specific intervention programs" (p 1176) and that "the success of specific intervention programs must be assessed according to specific criteria related to the content of the program" (p 1177).

Few early childhood programs seek to improve overall intelligence or to hasten the general cognitive development of infants and toddlers. Rather most programs seek to provide interventions for specific identified needs, either for the family or child or both. The typical early childhood program can be accurately viewed as a collection of individually tailored programs. Thus, the individual intended outcomes should be identified and the program's success gauged against whether those outcomes are worthwhile and whether they were attained.

The measures used to describe the development of program participants should not be accepted at face value. They are not necessarily reliable or valid for specific programs. Do the measures assess the relevant outcomes? Were they developed and normed on a population similar to that being served? If not, then a local item analysis and perhaps test recalibration is needed.

In lieu of control-comparison group hypothesis testing, we advocate the use of case studies, the computation of effect sizes, and the examination of growth curves. Case studies can provide rich data to help policy makers and researchers understand interventions. Effect sizes help gauge the relative contributions of the intervention. Growth curves can help identify trends and control for some error.

References

Administration for Children, Youth and Families (1983) The effects of the Head Start Program on children's cognitive development. ERIC Document No ED 248989.

Applegate, B. (1986) A meta-analysis of the effects of day-care on development: Preliminary Findings. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis, TN. ERIC Document No ED 280613.

Bayley, N. (1955) On the growth of intelligence, *American Psychologist*, 10, 805, Dec.

Bayley, N. (1958) Value and Limitations of infant testing, *Children*, 5 (4), 129-133.

Bayley, N. (1970) Development of mental abilities. In P.H. Mussen (ed) *Carmichael's manual of child psychology*, 1, New York: Wiley.

Escalona, S.K. & A. Moriarty (1961) Prediction of school age intelligence from infant tests. *Child Development*, 32, 597-605.

Goffeney, B, Henderson, N, & B. Butler (1971) Negro-white, male-female eight month developmental scores compared with seven year WISC and Bender test scores, *Child Development*, 42, 595-604

Goldring, E. & L. Presbrey (1986) Evaluating preschool programs: A meta-analytic approach, *Educational Evaluation and Policy Analysis*, 8(2) 179-188, Summer.

Horner, T.M. (1980) Test-retest and home-clinic characteristics of the Bayley Scales if Infant Development in Nine- and fifteen-month old infants, *Child Development*, 51, 761-758.

Lewis, M. & H. McGurk (1972) Evaluation of Intelligence, *Science*, 178, 1174-1177, Dec 15.

Ottenbacher, K.J. (1992), Practical significance in early intervention research: From affect to empirical effect. *Journal of Early Intervention*, 16 (2), 181-193, Spring.

Thomas, H. (1970) Psychological assessment instruments for use with human infants. *Merrill-Palmer Quarterly*, 16, 179-223.

Werner, E.E. & N. Bayley (1966) The reliability of Bayley's revised scale of mental and motor development during the first year of life. *Child Development*, 37, 39-50.

This publication was prepared with funding from the Office of Educational Research and Improvement, U.S. Department of Education, under contract RR93002002. The opinions expressed in this report do not necessarily reflect the positions or policies of OERI or the U.S. Department of Education. **Permission is granted to copy and distribute this ERIC/AE Digest.**